

Sample size calculations for impact evaluations

Marcos Vera-Hernandez

UCL, IFS, and PEPA

m.vera@ucl.ac.uk

Objectives

- When doing an impact evaluation, sometimes we can use existing data. Some other times, we need to collect our own data. Then:
 - What sample size should I use?
 - Considerations to minimize data collection costs
 - How should adjust the sample size computation if I use more than one outcome to assess the intervention
- Emphasis on practical issues rather than theoretical concepts

Topics to be covered

- Setting the stage/ basic definitions
- Sample size with continuous outcomes
 - Computations
 - Cost minimization
 - Data loss
- Sample size with proportions
 - Computations
- Cluster based impact evaluations
 - Partial take-up and non-compliance
 - Intra cluster correlation
 - Computations
 - Cost minimization
- Adjustment for multiple outcomes

Setting the stage

- An intervention to evaluate
 - Training program (six months starting July 2014)
- An outcome variable to assess the success of the intervention
 - Labour earnings as of December 2015
- Intervention was allocated using a randomized experiment:
 - Treatment and Control
 - Treated = participants, Control = non-participants

Setting the stage

- Method to estimate the effect of the intervention on the outcome variables (earnings)
 - We need to interview treated and control individuals in January 2016
 - There are many individuals
 - We will draw a **random sample** of them (and ask them for their earnings)

Setting the stage

- Compute **average earnings** for individuals who were allocated to **treatment**
- Compute **average earnings** for individuals who were allocated to **control**
- The difference between these two averages include a *random* component because we only used a *sample* to compute them (not the entire population)
- We carry out a test of hypothesis to see if the difference is **statistically significant**

Power and Significance

- Null hypotheses to test:
 - The intervention does **not** have an effect on mean earnings
 - The mean earnings of treated individuals is the same as the mean earnings of control individuals
 - This average is not the average in the sample drawn to do the interviews and gather the data, it is the average on the entire population
 - Note, the null hypothesis is **pessimistic**
 - The intervention will be deemed successful if the null hypothesis is **rejected**

Significance

- Probability that our test will **reject** the null hypothesis when it is **true**
 - Also known as α
- Probability that our test will say that the average earnings of treatment and control individuals is different when in reality they are the same
- Usually, we ask this probability to be 0.05 (or smaller)

Power

- Probability that our test will **reject** the null hypothesis when it is **false**
- Probability that our test will say that the average earnings of treatment and control individuals is different when in reality they are different
- Usually, we ask this probability to be **0.80** (or higher)

Conducting sample size calculations for continuous outcomes

What is the minimum sample size required to achieve an acceptable level of power (usually 0.8) at a given significance level (usually 0.05)?

The test comprises the difference in the average of the outcome variable between treated and control individuals (**test of means**)

We will focus on two-sided tests because they are the most common on impact evaluations

- Example of continuous outcomes: *earnings*
- What information is needed?
 - Power and α
 - Mean of outcome variable for treated (μ_A)
 - Mean of outcome variable for controls (μ_B)
 - Standard deviation of the outcome variables for treated and controls
 - Usually assumed to be the same
 - Ratio of sample in treated versus controls (n_A/n_B)
 - Usually 1 but there might be good reasons to have other ratios (see later on)

Basic free software:

<http://powerandsamplesize.com/Calculators/Compare-2-Means/2-Sample-Equality>

<http://www.stat.ubc.ca/~rollin/stats/ssize/n2.html>

<http://statpages.org/#Power>

Subroutines in STATA (i.e. sampsi) and R

An Example for continuous outcomes

Using available data sources (i.e. Labour Force Survey, British Household Panel Survey...) and/or other published studies, I estimate that:

The mean of monthly earnings will be **600** for the controls
($\mu_B=600$)

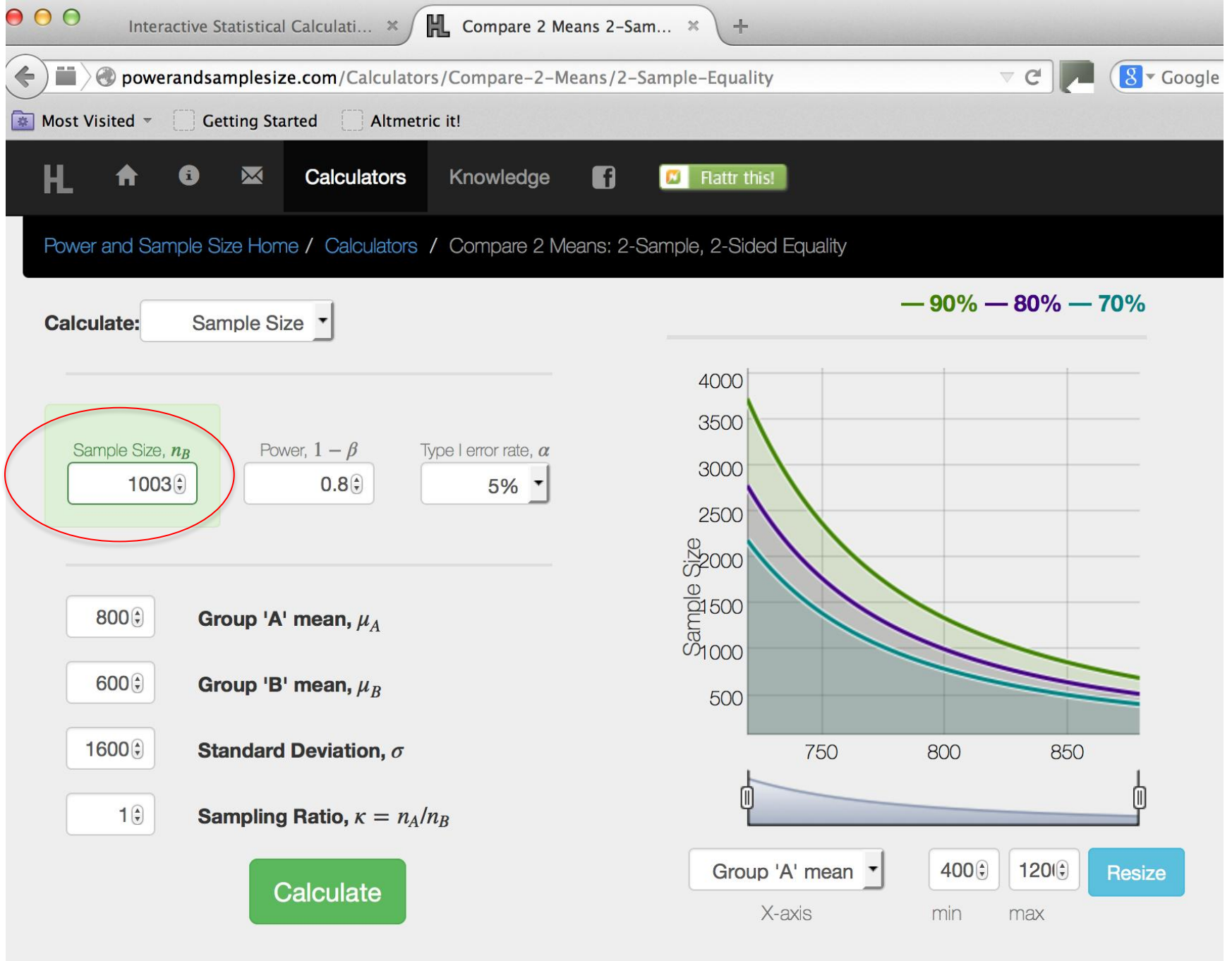
The standard deviation of monthly earnings will be **1600**: ($\sigma=1600$)

- I will assume the same standard deviation for treated and controls

Based on my reading of existing studies, I expect that the mean of monthly earnings will be **800** for the treated ($\mu_A=800$)

Other standard parameters: **Power =0.80**, **$\alpha=0.05$** , Ratio $n_A/n_B = 1$

<http://powerandsamplesize.com/Calculators/Compare-2-Means/2-Sample-Equality>



We will need data on 1003 individuals in the control group (**$n_B = 1003$**)

The software only give us the sample size in group B. However, as we specified a ratio of 1 (**$n_A/n_B = 1$**), this means that we will also need data on 1003 individuals of the control group (**$n_A = 1003$**)

Total sample size: $1003+1003 = \mathbf{2006}$

What happens if we specify a different ratio? For example: (**$n_A/n_B = 2$**)

Calculate:

Sample Size ▾

Sample Size, n_B

752

Power, $1 - \beta$

0.8

Type I error rate, α

5% ▾

800

Group 'A' mean, μ_A

600

Group 'B' mean, μ_B

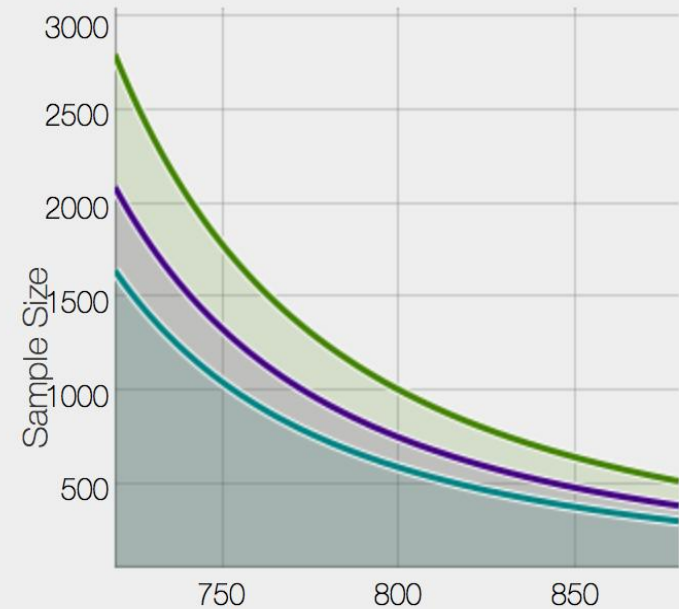
1600

Standard Deviation, σ

2

Sampling Ratio, $\kappa = n_A/n_B$

Calculate



Group 'A' mean ▾

400

1200

Resize

X-axis

min

max

What happens if we specify a different ratio?

For example: ($n_A/n_B = 2$)

In this case, $n_B = 752$

Using the relation ($n_A/n_B = 2$), we can easily obtain

$$n_A = 2 \times 752 = 1504$$

Note that **$1504+752=2256$** which is larger than **2006**. The **smallest total sample size is achieved with a ratio 1**

Some important relations

- The larger the power, the larger the sample
- The smaller α , the larger the sample
- The larger the difference between the means of the treatment and control group, the smaller the sample
- The larger the standard deviation, the larger the sample

Cost minimization

- Sometimes, it is much more costly to collect information from treated individuals than from controls
 - For instance, we might have to pay the training program for treated individuals
 - In this case, it makes sense to try several ratios to see which one minimizes the cost
 - Of course, the books also contain formulas that give you the result directly

Cost Minimization

| Ratio | N Controls | N treated | Cost per treated | Cost per control | Total cost |
|-------|------------|-----------|------------------|------------------|------------|
| 1 | 1003 | 1003 | 2000 | 1400 | 3410200 |
| 0.9 | 1059 | 953.1 | 2000 | 1400 | 3388800 |
| 0.8 | 1128 | 902.4 | 2000 | 1400 | 3384000 |
| 0.7 | 1218 | 852.6 | 2000 | 1400 | 3410400 |
| 0.6 | 1337 | 802.2 | 2000 | 1400 | 3476200 |

All these combinations give you the same power at the same α , so it makes sense to use the one that minimizes the cost

Data loss

- The sample size obtained are the minimum sample size to achieve the specified power at the specified α
- But it is realistic to think that some data will be lost:
 - An individual that agrees to participate but later on he/she changes his mind
 - We will not be able to locate an individual that initially agreed to participate
 - Some individuals will not want to answer the question on earnings
 - Some other data (gender, age, education) might be missing
- We should inflate the minimum sample size by some percentage to allow for this loss of data

Conducting sample size calculations for outcomes which are measured as proportions

What is the minimum sample size required to achieve an acceptable level of power (usually 0.8) at a given significance level (usually 0.05)?

We will focus on two-sided tests because they are the most common on impact evaluations

- Example: *proportion unemployed*
- What information is needed?
 - Power and α
 - Proportion for treated (p_A)
 - Proportion for controls (p_B)
 - The standard deviation is not needed for such data as proportions
 - Ratio of sample in treated versus controls (n_A/n_B)
 - Usually 1 but there might be good reasons to have other ratios (see later on)

An Example for proportions

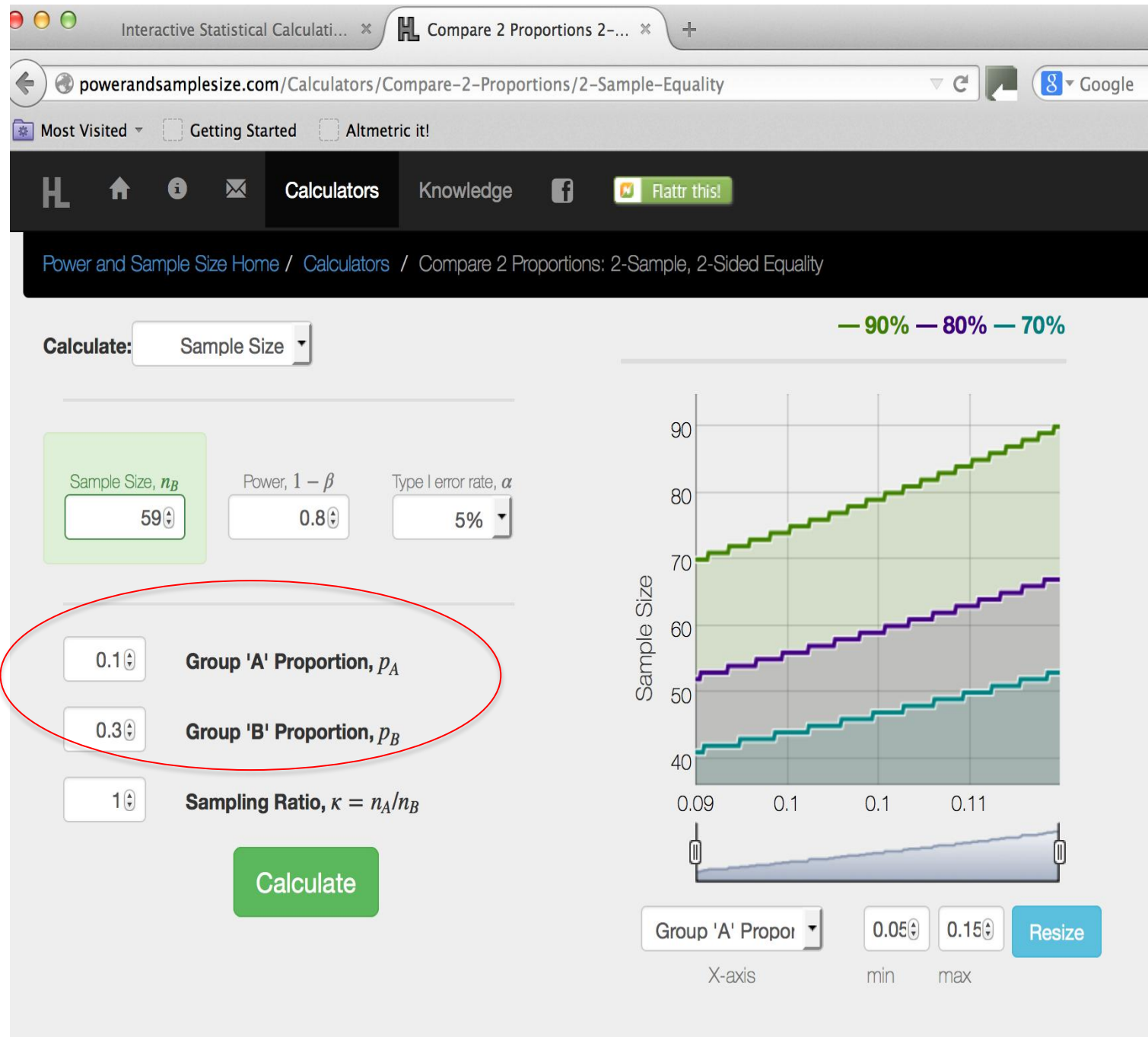
Using available data sources (i.e. Labour Force Survey, British Household Panel Survey...) and/or other published studies, I estimate that:

The proportion of unemployed control individuals will be 0.3 ($p_B=0.3$)

Based on my reading of existing studies, I expect proportion of unemployed treated individuals to be 0.1 ($p_A=0.1$)

Other standard parameters: **Power =0.80**, **$\alpha=0.05$** , Ratio $n_A/n_B = 1$

<http://powerandsamplesize.com/Calculators/Compare-2-Proportions/2-Sample-Equality>



About proportions

- For a given difference between p_A and p_B , the closer the proportions are to 0.5, the larger the sample required will be
 - This is because the standard deviation is higher when the proportions are close to 0.5
- All other considerations are the same as in the continuous outcomes, including data loss, cost minimization...

Cluster based Impact Evaluations

Cluster based impact evaluations

- Outcomes at the individual level (earnings, test scores...) but intervention is allocated at a higher level (area, school,...)
 - Because of logistical/political feasibility or to avoid contamination from treatment to control (i.e. two students in the same class)
 - The higher level (school, area...) is referred as the cluster
- Computations have to be adjusted for two factors
 - Partial take-up and non-compliance
 - Statistical dependence of individuals belonging to the same cluster

Partial take-up and non-compliance

- The training intervention that we are evaluating is for individuals aged 18-20 and takes place in some local area authorities (LEA) and not others
 - The cluster is the LEA
 - Only 80% of individuals aged 18-20 sign up for the intervention in treated LEA
 - 10% of individuals aged 18-20 sign up for the intervention in control LEA (administrative error...)
- We must re-adjust the income averages of 600 & 800
 - $\mu_A = 0.8 \times 800 + (1 - 0.8) \times 600 = 760$
 - $\mu_B = 0.9 \times 600 + (1 - 0.9) \times 800 = 620$
- You would do the same for proportions
- Sometimes, you do not need to adjust because take-up is full and non-compliance is null

Statistical dependence within cluster

- Things happen within a cluster (school, area...) that make individuals to have similar values of the outcome variable (teacher quality, plant closures,...)
 - If I already collected data from 10 people from the same area, collecting data from another person from the same area does not add much additional information (if the statistical dependence is strong)
- The statistic to measure such statistical dependence is the *Intra Cluster Correlation*

Statistical dependence within cluster

- *Intra Cluster Correlation (ICC):*
$$\rho = \frac{S_b^2}{S_b^2 + S_w^2}$$
- S_b^2 = variance of the outcome variable (i.e. earnings) between clusters
 - Compute the average of the outcome variable in each cluster
 - Compute the variance of those averages
- S_w^2 = variance of the outcome variable (i.e. earnings) within clusters
 - Compute the variance outcome variable in each cluster
 - Compute the average of all those variances
- In Stata, you can easily compute it using: *loneway earnings idcluster*
- If there is maximum dependence (all individuals in the same cluster have the same value of the outcome variable) then $S_w^2=0$, then the **ICC** is **1**
- If there is no dependence within clusters (all clusters have the same average value of the outcome variable) then $S_b^2=0$, then the **ICC** is **0**
- The ICC is between 0 and 1. Usually, the larger the cluster the smaller the ICC is

Statistical dependence within cluster

- *Intra Cluster Correlation (ICC):*
$$\rho = \frac{S_b^2}{S_b^2 + S_w^2}$$
- How do I know that ICC for my sample size calculation?
 - You can either use a dataset similar to the one that you will be collecting (same outcome variable, same definition of cluster, similar population)
 - You can obtain it from published articles or from publicly accessible research proposals

Statistical dependence within cluster

- *Design Effect (DE):*

$$DE = 1 + (m - 1)\rho$$

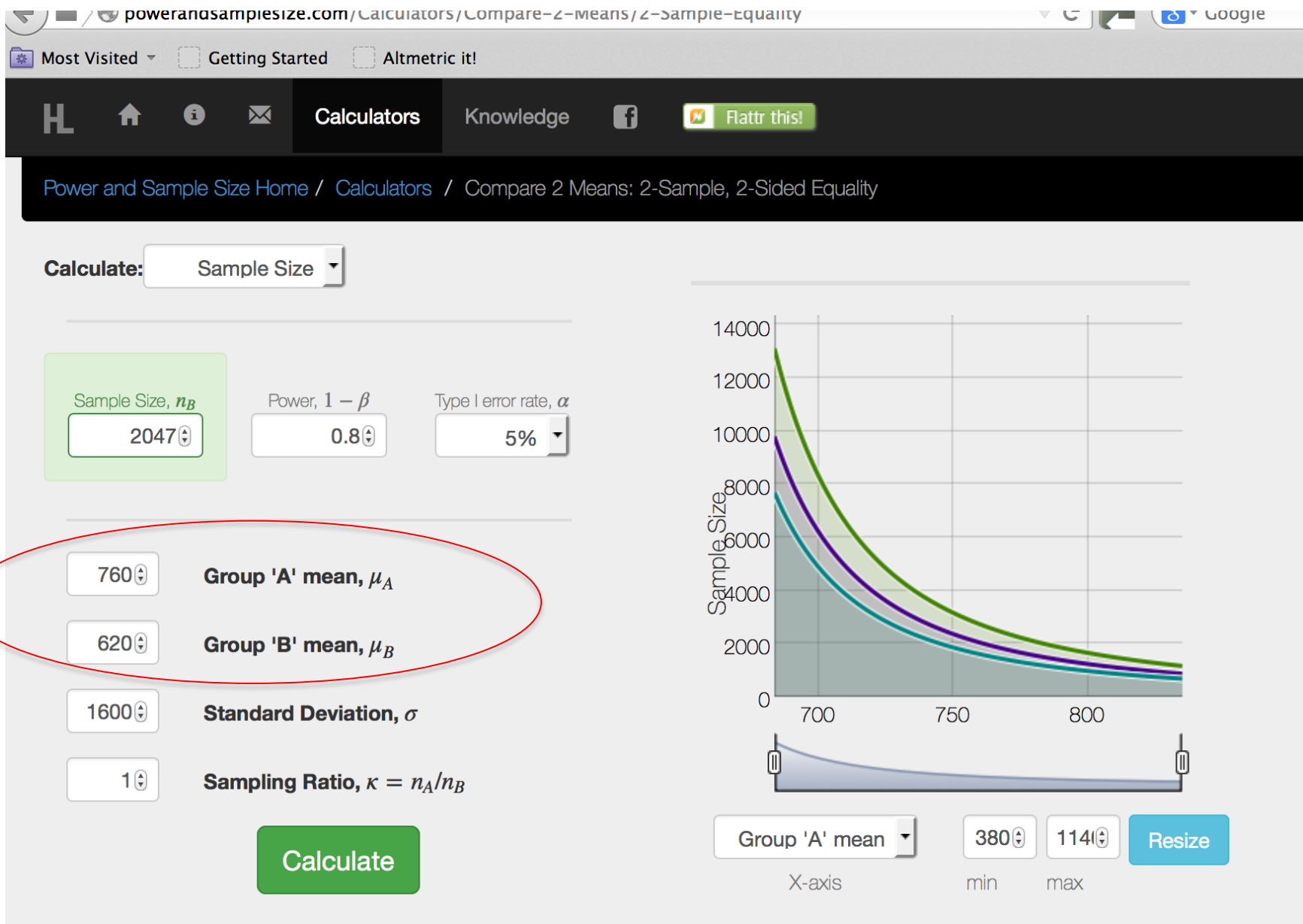
m = number of individuals per cluster

ρ = intra cluster correlation coefficient

Once we have computed the sample size ignoring any statistical dependence within cluster (using the methods previously outlined), then we must multiply that sample size by the **Design Effect** to obtain the total sample size adjusted for statistical dependence within clusters

Example cluster based impact evaluation

- Once we have adjusted for partial take-up and non-compliance, we assume that:
 - $\mu_A = 0.8 \times 800 + (1 - 0.8) \times 600 = 760$
 - $\mu_B = 0.9 \times 600 + (1 - 0.9) \times 800 = 620$
- We use software to compute the required sample size ignoring any issues related to statistical dependence within clusters
- We obtain $2047 \times 2 = \mathbf{4094}$ individuals using standard levels of power and α (see screenshot in following slide)



Example cluster based impact evaluation

- Now, we must take the 4094 and adjust it for statistical dependence
- Using published studies, we assume that the ICC of earnings (at the area level) will be 0.05
 - If we have 10 individuals per cluster, the DE will be $1+(10-1) \times 0.05 = 1.45$
 - If we have 15 individuals per cluster, the DE will be $1+(15-1) \times 0.05 = 1.70$
- If we choose 10 individuals per cluster, we multiply 4094 by 1.45 = 5936 total number of individuals (the total number of clusters will be 5936/10=594)
- If we choose 15 individuals per cluster, we multiply 4094 by 1.70 = 6960 total number of individuals (the total number of clusters will be 6960/15=464)

Software for cluster based IE

- You do not need special software, you can use what we have just done in a spreadsheet
- The software below will not adjust for the partial take-up or non-compliance directly, you have to make the adjustment before entering the parameters
- STATA:
 - Install the user command *sampclus*
 - Type: *sampsi: 760 620, sd(1600) power(0.8)*
 - And then type: *sampclus, rho(0.05) obsclus(10)*
 - An alternative is to use this other STATA command that allows to enter different number of treated and control clusters:
 - <http://www.population-health.manchester.ac.uk/biostatistics/research/software/clsampsi/>
 - This can be very useful if the cost of a treatment cluster is different than the cost of a control cluster or if there are other constraints that limit the number of either treatment and control clusters
- *Optimal Design:*
 - <http://hlmssoft.net/od/>
 - http://sitemaker.umich.edu/group-based/optimal_design_software
 - For proportions, this software uses a different way of adjustment

Cost Minimization for cluster based IE

- As we saw, there are different combinations of number of clusters and number of people per cluster that give the same power
 - If we choose 10 individuals per cluster, we need 5936 individuals (594 clusters, each with 10 individuals)
 - If we choose 15 individuals per cluster, we need 6960 individuals (464 clusters, each with 15 individuals)
- Usually there is a fixed cost of going into a cluster plus a marginal cost of interviewing a person within each cluster
- By trying different number of individuals per cluster, we can choose the combination of number of clusters and number of individuals per cluster that minimizes the cost
- Software such as *clamps* or *Optimal Design* can also do this for us

Cost Minimization for cluster based IE

- If the cost of each treatment cluster is higher than the cost of each control cluster, we might be able to further decrease costs (but keep the same level of power) by using more control clusters than treatment clusters
 - Software *clsampsi* might be very useful for this (<http://www.population-health.manchester.ac.uk/biostatistics/research/software/clsampsi/>)

Correction for Multiple Outcomes

Multiple Outcomes

- It is standard to use more than one indicator to assess the success of the intervention
- For instance: earnings and unemployment
- We want to be sure that the probability of rejecting any of the null hypotheses (two in this case) when they are both true is still 0.05
 - This is known as the Familywise Error Rate
 - This will not be true if we use an α of 0.05 for **each** hypothesis
 - *Probability* of rejecting either earnings or unemployment =
 - $1 - \text{Probability of not rejecting either earnings or unemployment}$
 - $1 - (1 - 0.05)(1 - 0.05) = 0.0975$ which is larger than 0.05
 - The intuition is that because we are carrying out more than one test, it gets easier to reject one of them (because of bad luck) even if they are all true
- This will mean that we will have to use an α for each test that is smaller than 0.05 (adjustment)

Adjustments for multiple outcomes

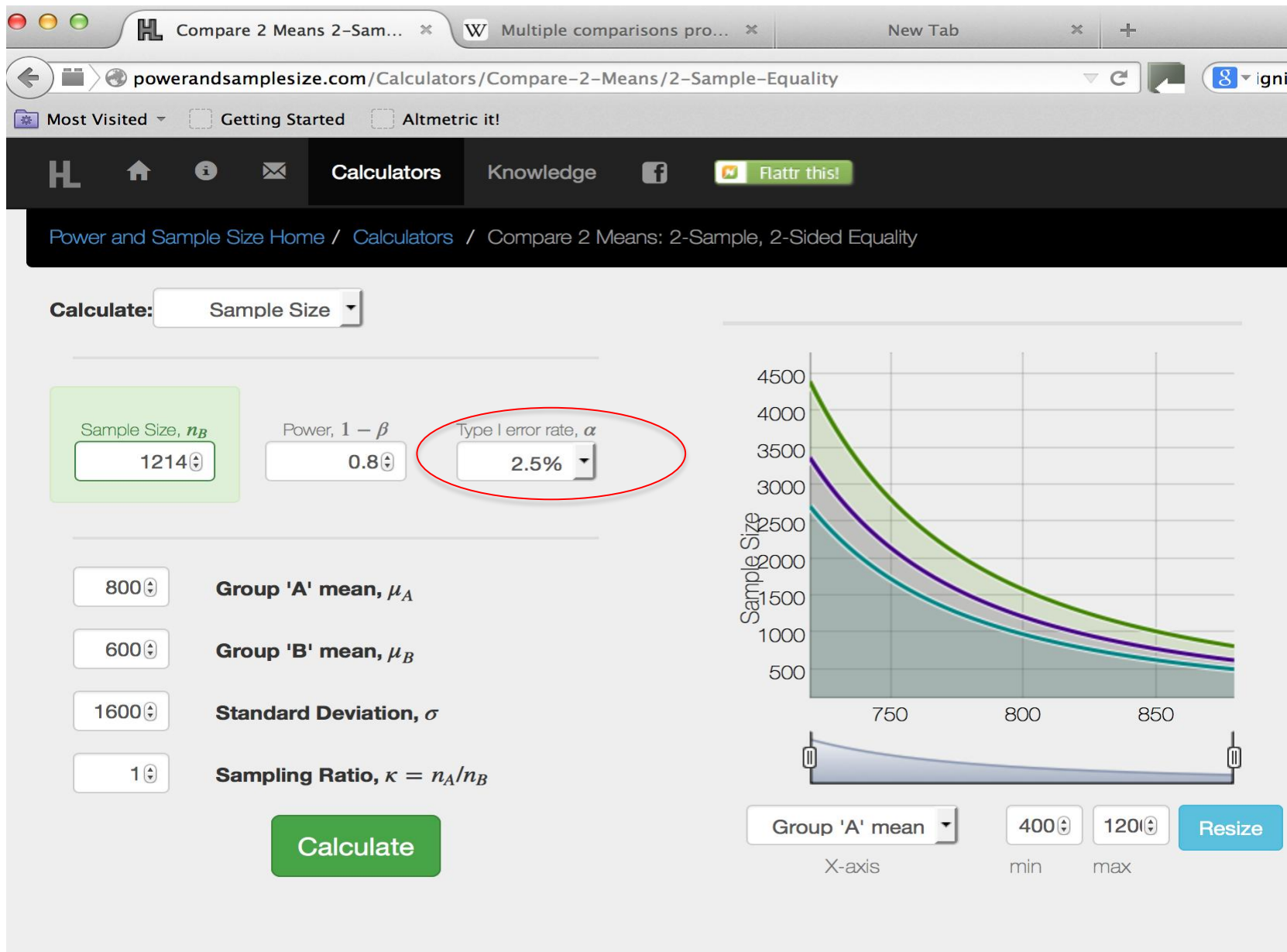
- They consist of using an α for each test that is smaller than 0.05, so that the FWER is kept at 0.05
- Many methods:
- Bonferroni:
 - $\alpha = \text{FWER} / (\text{Number of Outcomes})$
 - $\alpha = 0.05/2 = 0.025$
- Tukey:

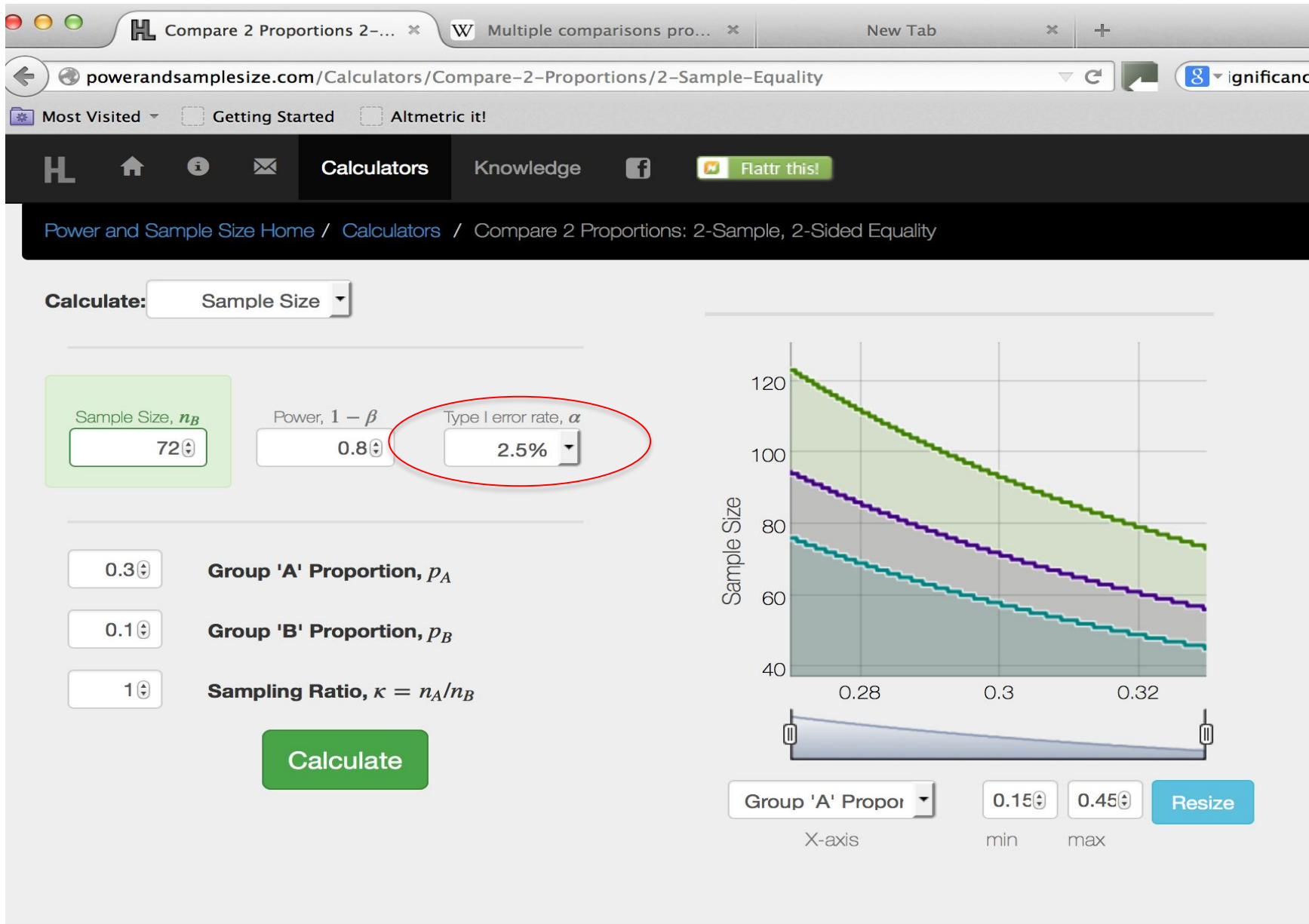
$$\alpha = 1 - (1 - \text{FWER})^{1/\sqrt{\text{number outcomes}}}$$

$$\alpha = 1 - (1 - 0.05)^{1/\sqrt{2}} = 0.0356$$

Example of adjustments for multiple outcomes

- We go back to the two examples of the beginning: earnings and proportion of unemployed
- But we carry out the computations using an α of **0.025** (following the Bonferroni method) so that the FWER is 0.05





Example of adjustments for multiple outcomes

- So we will need $1214 \times 2 = 2428$ individuals for earnings and $72 \times 2 = 144$ for proportion of unemployment
- Clearly, we must go for the maximum of these two numbers: 2428 individuals
- Note that this is larger than the number that we computed at the beginning (2006) that did not adjust for multiple outcomes

Thank you!